RECEIVED
CENTRAL FAX CENTER

JAN 17 2006

PATENT

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re Application of:
Reiner Kraft et al.
Serial No.: 09/607,370
Group Art Unit: 2178
Filed: June 30, 2000
Examiner: Kyle R. Stork
For: *SYSTEM AND METHOD FOR ENHANCED BROWSER-BASED WEB CRAWLING*

APPEAL BRIEF

VIA FACSIMILE (571) 273-8300
MS-APPEAL BRIEF-PATENTS
Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

Sir:

This Appellants' Amended Brief is filed in response to a Final Office Action dated July 15, 2005, an Advisory Action dated September 20, 2005, and a Notice of Appeal received October 17, 2005, the due date to which, has been extended to January 17, 2005 by the enclosed petition for extension of time. Reconsideration of the Application, withdrawal of the rejections, and allowance of the claims are respectfully requested.

CERTIFICATE OF MAILING/FACSIMILE

I hereby certify that this correspondence is being deposited with the United States Postal Service as first class mail in an envelope addressed to: Mail Stop Appeal Brief-Patent, Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450 or facsimile transmitted to the U.S. Patent and Trademark Office on the date shown below.

ON: 1/17/06 (DATE) BY: Karen Taragowski

Karen Taragowski

I. REAL PARTY IN INTEREST

The real party in interest is International Business Machines (IBM) of Armonk, NY.

II. RELATED APPEALS AND INTERFERENCES

There are no related appeals or interferences.

III. STATUS OF CLAIMS

Claims 1-20 are pending.

Claims 1 through 20 are rejected.

The Appellants are appealing the rejection of independent claims 1, 14, and 20 (all other remaining claims depend from these claims). Claims 1, 14, and 20 are on appeal.

IV. STATUS OF AMENDMENTS

The Examiner issued a final rejection of claims 1-20 in the Final Office Action of July 15, 2005. Appellants submitted a response without amendment to this Final Office action to overcome the Examiner's rejections. The Advisory Action dated September 20, 2005 addressed the Appellants' remarks and indicated that the response without amendment was entered.

V. SUMMARY OF THE CLAIMED SUBJECT MATTER

This summary references line numbers of the specification as filed. It is to be noted that the text of each page of the filed specification starts with line number 5.

As set forth by the independent claims on appeal, the claimed subject matter includes a retrieval unit for retrieving a web document at an address. FIG. 3a: reference 308a; FIG. 3b: reference 302b; FIG. 5: references 504 and 506; and Specification at page 10, lines 14-15 and page 12, lines 1-3. The retrieval unit is also for extracting contents of the web document. FIG. 3a: reference 310a; and Specification at page 12, lines 11-13. The retrieval unit retrieves the

document and extracts its contents for rendering an intermediate dynamically constructed in-memory webpage representation of the web document at a hub processing unit. FIG. 3a: reference 310a; FIG. 6 generally; and Specification at page 10, lines 17-19 and at page 12, lines 8-29 to page 13, line 1. The intermediate dynamically constructed in-memory webpage representation is formatted as if displayed for viewing on an end-user's web browser. See Id.

The claimed subject matter further includes a loader for loading secondary documents as required associated with the web document in order to render the secondary documents as part of the in-memory webpage representation. FIG. 3b: reference 304b; FIG. 6: references 606-614; and Specification at page 10, lines 15-17 and page 13, lines 2-5. The secondary documents include one or more images with textual content embedded therein. See Id. The hub processing unit renders the in-memory webpage prior to analyzing and summarizing the in-memory webpage. FIG 3b: references 306b, 308b, 310b; and Specification page 12, lines 8-29 to page 13, lines 1-8 and page 14, lines 1-16.

The claimed subject matter also includes a summarizer (reference 316a). for analyzing and summarizing the in-memory webpage representation. Fig. 3a: reference 314a; FIG 3b: reference 310b; FIG. 7 and FIG. 8 generally; and Specification at page 10, lines 19-22 and page 14, lines 2-16. The summarizer (reference 316a) produces a text map for the webpage document of the textual contents therein. FIG. 7 generally; and Specification at page 13, lines 18-29.

The claimed subject matter also includes an optical character recognition engine for use on the images to extract textual content for adding to the textual map for the webpage document. FIG. 7: reference 712; and Specification at page 13, lines 25-28.

The present invention is advantageous over the prior art for many reasons. First, the present invention permits the fault-tolerant gathering of dynamic data documents on the World Wide Web. For example, the present invention is able to summarize web documents containing executable client side software code. Second, the present invention allows for the interpretation and

summarization of textual and other information contained within the body of a web-based image document. In one embodiment, the present invention implements optical character recognition with web crawling so that a web crawler is able to properly summarize images and image maps that contain textual or other data.

The Enhanced Browser Based Crawler of the present invention enhances existing document gathering and analysis by, for example, dramatically improving the quality of the extracted metadata. This is due to the fact that the summarization of a document is based on the whole and complete document as it was designed by the document's author; the static heterogeneous data as well as the problematic dynamic data is completely rendered and integrated into the metadata for subsequent indexing of all metadata by a web crawler. For example, a dynamic in-memory representation of the web page, as intended to be seen by an end user, is created to extract the most accurate and comprehensive data set possible. A standard web crawler is not able to compose this type of highly dynamic and distributed document that includes dynamic information such as client side script, applets, or their equivalents.

VI. GROUNDS OF REJECTION TO BE REVIEWED ON APPEAL

Whether claims 1, 14, and 20 are unpatentable under 35 U.S.C. §103(a) over *Meyerzon et al.* (U. S. Patent No. 6,638,314) in view of *Lawrence et al.* (U.S. Patent No. 6,026,409) and in further view of *Blumenthal* (U. S. Patent No. 6,638,314) .

VII. ARGUMENT

A. WHETHER CLAIMS 1, 14, AND 20 ARE UNPATENTABLE OVER *MEYERZON* IN VIEW OF *LAWRENCE AND IN FURTHER VIEW OF BLUMENTHAL*

In the Examiner's Office Action of July 15, 2005, the Examiner rejected claims 1-3, 14-16, and 20 under 35 U.S.C. §103(a) as being unpatentable over Meyerzon et al. (U.S. Patent No. 6,638,314) (Hereinafter *Meyerzon*), in view of Lawrence et al. (U.S. Patent No. 6,289,342) (Hereinafter *Lawrence*) and in further view of Blumenthal (U.S. Patent No. 6,026,409) (Hereinafter *Blumenthal*). The Appellants respectfully submit that claims 1-20 are patentable over *Meyerzon* and/or *Lawrence* and/or *Blumenthal* under 35 U.S.C. § 103(a). The Appellants assert that neither the *Meyerzon, Lawrence*, or *Blumenthal* references, taken either alone or in combination with one another, teach or suggest the claimed limitations, particularly the claim limitation of "rendering an intermediate dynamically constructed in-memory webpage representation of the web document at a hub processing unit which is formatted as if displayed for viewing on an end-user's web browser."

Appellants respectfully suggest selection of independent claim 1 as representative of the independent claims on appeal. Independent claim 1 is directed towards a method for browser-enhanced web crawling associated with a network of hub processing units coupled to a plurality of information processing units over a network, the method executed by a web crawler on a hub processing unit associated with the network comprising:

> retrieving a web document at an address, and extracting contents of the web document for rendering an intermediate dynamically constructed in-memory webpage representation of the web document at a hub processing unit which is formatted as if displayed for viewing on an end-user's web browser;
>
> loading secondary documents associated with the web document in order to render the secondary documents as part of the in-memory webpage representation, wherein the secondary documents include one or more images with textual content embedded therein, wherein the hub processing unit renders the in-memory webpage prior to analyzing and summarizing the in-memory webpage;

analyzing and summarizing the in-memory webpage representation to produce a text map for the webpage document of the textual contents; and

using optical character recognition on the images to extract textual content for adding to the textual map for the webpage document.

The Appellants assert that, in particular, the underlined portions of the above claims are not taught or suggested by the *Meyerzon* and/or *Lawrence* and/or *Blumenthal* references, taken either alone or in combination with one another.

The claims were rejected under 35 U.S.C. §103(a). The Statute expressly requires that obviousness or non-obviousness be determined for the claimed subject matter "as a whole," and the key to proper determination of the differences between the prior art and the present invention is giving full recognition to the invention "as a whole." As discussed below, the Appellants assert that these limitations, especially when considered in the context of the other limitations of claim 1, are not described in the prior art references of record and that these limitations render the claimed subject matter unobvious over the prior art.

Overview of Prior Art

To begin, the *Meyerzon* reference is directed towards a web crawler program that includes a gatherer process for gathering information pertaining to electronic documents. See *Meyerzon* at col. 8, lines 58-60. In the system of *Meyerzon*, worker threads process URLs and then pass each URL to a filter daemon. See *Meyerzon* at col. 9, lines 13-16. The filter daemon uses the URL to retrieve the electronic document at the address specified by the URL. See *Meyerzon* at col. 9, lines 16-20. After retrieving an electronic, the filter daemon parses the electronic document and returns a list of text and properties. See

Meyerzon at col. 9, lines 29-31. The worker thread then passes the list of properties and text to the indexing engine for creating an index which is used by the search engine in subsequent searches. See *Meyerzon* at col. 10, lines 13-16. A user may then examine the list of documents returned by the search engine, select a document and then the web browser displays the selected document to the user. See *Meyerzon* at col. 8, lines 23-25 and 32-35.

The *Lawrence* reference is directed to an autonomous citation indexing system that can be used as an assistant agent for automating and enhancing the task of finding publications in electronic form such has on the world wide web. See *Lawrence* at the Abstract. A citation index is autonomously created from literature in electronic form by an autonomous citation index ("ACI"). See *Lawrence* at column 7, lines 50-52. If the literature is not in electronic form, optical character recognition ("OCR") can be used to convert the literature to electronic form. See *Lawrence* at column 7, lines 51-53. The ACI system can then autonomously locate new articles, extract citations, identify citations to the same article which occur in different formats, and identify the context of citations in the body of articles. See *Lawrence* at column 7, lines 53-58.

The *Blumenthal* reference is directed towards a system and method for the visual search and retrieval of digital information within a single document of multiple documents. See *Blumenthal* at column 7, lines 9-11. A viewing window has a first pane that provides a global view of digitally stored information and a second pane that provides a local view of the information. A user submits queries and the keywords entered are displayed on the user's screen in a particular document as being highlighted. See *Blumenthal* at column 7, lines 1-25.

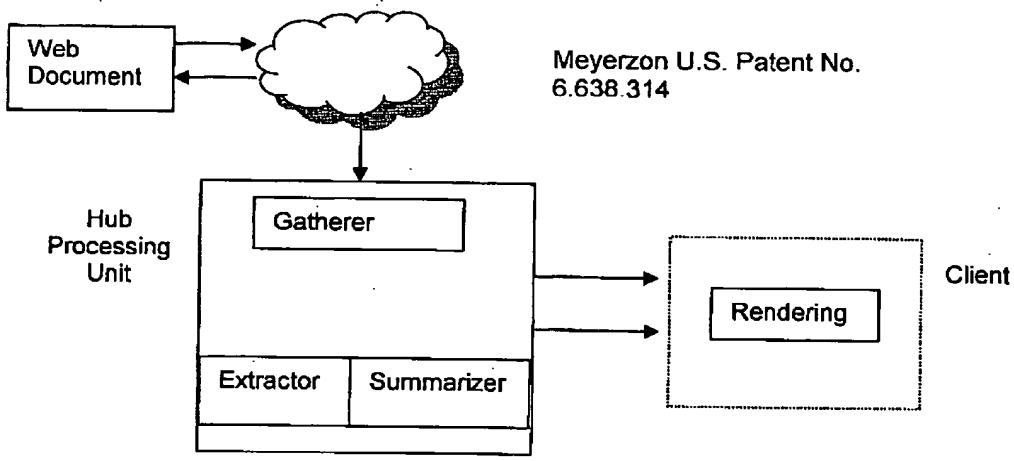
Cited References Fail to Describe All Limitations

With regards to the first limitation of claim 1, the Appellants traverse the Examiner's assertion that the *Meyerzon* reference discloses that "extracting contents of the web document for rendering an intermediate dynamically constructed in-memory webpage representation of the web document at a hub

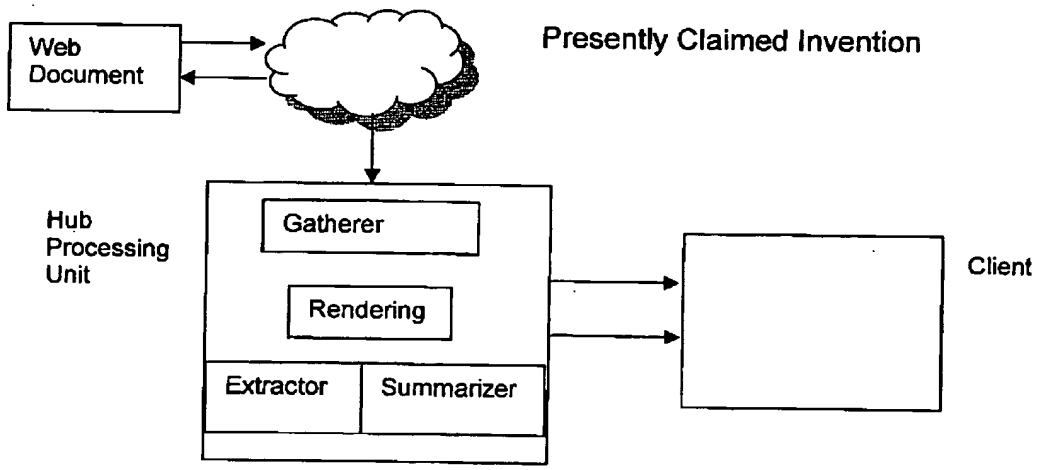
processing unit which is formatted as if displayed for viewing on an end-user's web browser". The Examiner, in the Final Office Action, cites *Meyerzon*, column 7, lines 60-65 and column 8, lines 15-10. The cited portions of the *Meyerzon* reference are limited to 1.) a web crawler retrieving electronic documents and data associated with the documents and 2.) a browser that locates and displays documents to a user. The Examiner, in the Advisory Action, further cites *Meyerzon*, column 2, lines 46-55; column 9, lines 29-59; column 11, lines 27-12 and 53; and column 14, lines 32-54. These cited portions of the *Meyerzon* reference are limited to 1.) retrieving a copy of the document and parsing the retrieved copy (column 2, lines 46-55); 2.) parsing the document to return a list of text and properties, the text and properties are obtained from tags within the html document (column 9, lines 29-59); 3.) various types of web crawls, i.e. "first full crawl", "full crawl", and "incremental crawl" (column 11, lines 27-12 and 53); and 4.) computing a hash value for the retrieved document (column 14, lines 32-54).

The Appellants respectfully assert that the retrieving and parsing of documents disclosed by the *Meyerzon* reference is not "retrieving a web document at an address, and extracting contents of the web document for rendering an intermediate dynamically constructed in-memory webpage representation of the web document at a hub processing unit which is formatted as if displayed for viewing on an end-user's web browser" as is recited for independent claims 1, 14, and 20. *Meyerzon* explicitly states that text and properties are obtained from tags within the HTML documents. See *Meyerzon* at column 9, lines 9-43. Therefore, *Meyerzon* is working on HTML source code, as compared to an "intermediate dynamically constructed in-memory webpage representation of the web document at a hub processing unit which is formatted as if displayed for viewing on an end-user's web browser", as recited for independent claims 1, 14, and 20.

For example, the following diagram illustrates how the *Meyerzon* reference renders the webpage at the client side only:



In contrast, the following diagram illustrates how the presently claimed invention renders an in-memory representation of the webpage:



As can be seen from the above diagrams, *Meyerzon* does not teach, suggest, or anticipate rendering an intermediate dynamically constructed in-memory webpage representation of the web document at a hub processing unit which is formatted as if displayed for viewing on an end-user's web browser. As stated above, the *Meyerzon* reference teaches indexing electronic documents. The web crawler program of *Meyerzon* retrieves electronic documents and associated data. See *Meyerzon* at col. 7, lines 60-67. The information is passed to an indexing engine which creates an index of the retrieved data. The index contains reference information and pointers to corresponding electronic documents, for example, keywords. See *Meyerzon* at col. 8, lines 1-16.

When a user requests a search, the search engines examines its index and returns a list of documents to the browser of the user's computer. See *Meyerzon* at col. 8, lines 26-35. The *Meyerzon* reference is not teaching extracting contents of the web document for rendering an intermediate dynamically constructed in-memory webpage representation of the web document at a hub processing unit which is formatted as if displayed for viewing on an end-user's web browser. In fact, the *Meyerzon* reference is teaching indexing information (i.e. extracting keywords) on electronic documents, which is not the same as extracting contents for rendering an intermediate dynamically constructed in-memory webpage representation of the web document. The *Meyerzon* reference especially does not teach rendering, by a hub processing unit, an in-memory webpage as if displayed for viewing on an end-user's web browser. The Examiner's citations of the *Meyerzon* reference and the remainder of the reference are completely absent a teaching of the above claim element. The advantage of the present of the summarization of a document being based on the whole and complete document as it was designed by the document's author; the static heterogeneous data as well as the problematic dynamic data is completely rendered and integrated into the metadata for subsequent indexing of all metadata by a web crawler is not realized by *Meyerzon*.

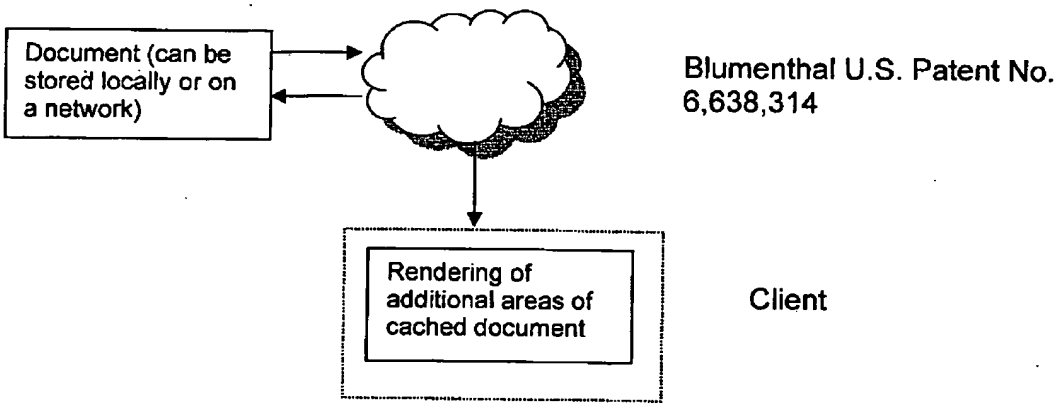
With regards to the second and third limitations of claim 1, the above arguments with respect to the "in-memory webpage representation of the web

document" are likewise applicable here and will not be repeated. As stated above, nowhere does the *Meyerzon* reference teach, anticipate, or suggest an "in-memory webpage representation of the web document" and therefore cannot teach, anticipate, or suggest "loading secondary documents associated with the web document in order to render the secondary documents as part of the in-memory webpage representation..." and/or "analyzing and summarizing the in-memory webpage representation..."

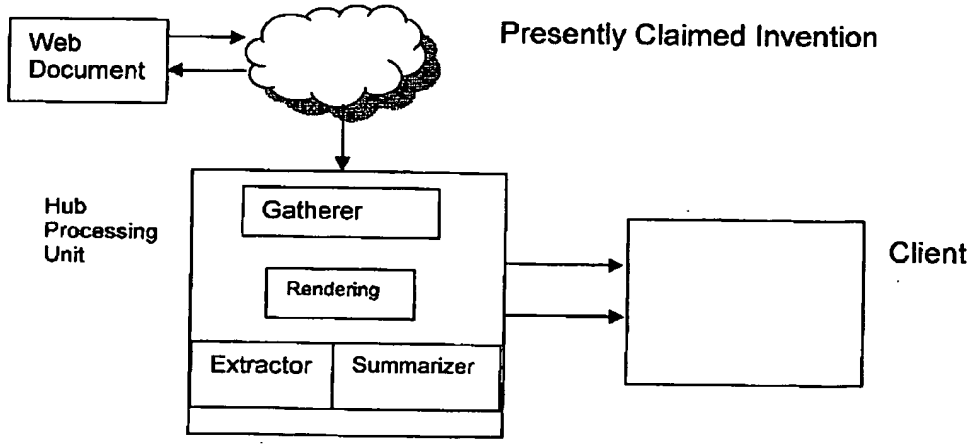
The Appellants further assert that the *Lawrence* reference is completely silent on "rendering an intermediate dynamically constructed in-memory webpage representation of the web document at a hub processing unit which is formatted as if displayed for viewing on an end-user's web browser". The *Lawrence* reference also does not teach, suggest, or anticipate loading secondary documents associated with the web document in order to render the secondary documents as part of the in-memory webpage representation, wherein the secondary documents include one or more images with textual content embedded therein, wherein the hub processing unit renders the in-memory webpage prior to analyzing and summarizing the in-memory webpage. The advantage of the present of the summarization of a document being based on the whole and complete document as it was designed by the document's author; the static heterogeneous data as well as the problematic dynamic data is completely rendered and integrated into the metadata for subsequent indexing of all metadata by a web crawler is not realized by *Lawrence*.

With regards to the *Blumenthal* reference, the Appellants traverse the Examiner's assertion that the *Blumenthal* discloses "wherein the hub processing unit renders the in-memory webpage prior to analyzing and summarizing the in-memory webpage". The Examiner, in the Final Office Action, cites *Blumenthal*, column 17, lines 45-53. The cited portions of the *Blumenthal* reference are limited to the additional rendering of a cached document at the client computer. For example, the following diagrams are provided to assist in describing the above technical differences between the *Blumenthal* reference and the present invention.

Starting with the *Blumenthal* reference, the following diagram illustrates how the additional rendering of areas of a cached document is rendered at the client side only:



In contrast, the following diagram illustrates how the presently claimed invention renders a complete in-memory representation of the webpage <u>at a hub processing unit</u>:



As can be seen from the above diagrams, the *Blumenthal* reference renders additional areas of the cached document on the client side. The present invention, on the other hand, renders the in-memory representation at a hub processing unit, as it would be displayed on a user's web browser.

Furthermore, the *Blumenthal* reference only teaches rendering additional areas of a cached document and not a complete in-memory webpage. See *Blumenthal* at col. 17, lines 45-52 and FIG. 13. The present invention on the other hand, renders an in-memory webpage as it would be displayed on a user's web browser and not just areas of the webpage. Therefore the present invention is able to summarize the document based on the whole and complete document as it was designed by the document's author; the static heterogeneous data as well as the problematic dynamic data is completely rendered and integrated into the metadata for subsequent indexing of all metadata by a web crawler. The *Lawrence* reference does not provide such an advantageous system.

Moreover, *Blumenthal* states at col. 17, lines 45-53 "...the cached document can be rendered..." wherein the term "render" relates to the visual display of a document and not the construction of an in-memory data structure, as recited for the present invention. See *Blumenthal* at col. 17, lines 45-53. The present invention, on the other hand, recites "renders the in-memory webpage" wherein the term "renders" is not implying a visual display of a document, but rather the construction of a data structure of the webpage in memory, which is subsequently analyzed and summarized. This distinction is important. The teachings of the *Blumenthal* reference are directed to the visual display of a document on a client. This is not an intermediary representation of the complete web page along with "the secondary documents" which are loaded "as part of the in-memory representation." The visual representation as described by the *Blumenthal* reference is not subsequently indexed. Accordingly, the teachings of the *Blumenthal* reference are completely *inoperable* in this regard.

Furthermore, as the references do not teach a hub processing unit for "retrieving a web document at an address, and extracting contents of the web document for rendering an intermediate dynamically constructed in-memory

webpage representation of the web document at a hub processing unit which is formatted as if displayed for viewing on an end-user's web browser... wherein the hub processing unit renders the in-memory webpage prior to analyzing and summarizing the in-memory webpage..." the Appellants respectfully assert that the suggestion for these elements cannot come from the Applicant's own specification. The Federal Circuit has repeatedly warned against using the Applicant's disclosure as a blueprint to reconstruct the claimed invention out of isolated teachings of the prior art. See MPEP §2143 and *Grain Processing Corp. v. American Maize-Products,* 840 F.2d 902, 907, 5 USPQ2d 1788 1792 (Fed. Cir. 1988) and *In re Fitch*, 972 F.2d 160, 12 USPQ2d 1780, 1783-84 (Fed. Cir. 1992). The references of *Meyerzon* and/or *Lawrence* and/or *Blumenthal* do not even suggest, teach or mention these claim limitations.

The references of *Meyerzon* and/or *Lawrence* and/or *Blumenthal* do not even suggest, teach or mention these claim limitations.

Moreover, the Federal Circuit has consistently held that when a §103 rejection is based upon a modification of a reference that destroys the intent, purpose or function of the invention disclosed in the reference, such a proposed modification is not proper and the *prima facie* case of obviousness can not be properly made. See In re Gordon, 733 F.2d 900, 221 USPQ 1125 (Fed. Cir. 1984). Here the intent, purpose and function of *Meyerzon* taken alone and/or in view of *Lawrence* and/or in further view of *Blumenthal* is the indexing of electronic documents for use by a search engine allowing a user to select a document to be displayed by a client-side web browser. The rendering of a webpage only occurs at the client side. Because *Meyerzon* does not render an in-memory webpage as it would be displayed on a user's web browser or render the in-memory webpage prior to analyzing and summarizing the in-memory webpage, this combination as suggested by the Examiner destroys the intent and purpose of *Meyerzon*. In contrast, the intent and purpose of the present invention is to render an in-memory webpage representation of a web document prior to analyzing and summarizing the in-memory webpage. Accordingly, the combination of *Meyerzon* and *Lawrence* in further view of *Blumenthal* results in

an inoperable system, and the Examiner's case of *"Prima Facie Obviousness"* should be withdrawn.

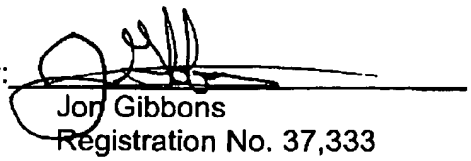
Additionally, the Federal Circuit stated in McGinley v. Franklin Sports, Inc., (Fed Cir 2001) that if references taken in combination would produce a "seemingly inoperative device," such references teach away from the combination and thus cannot serve as predicates for a prima facie case of obviousness. In re Sponnoble, 405 F.2d 578, 587, 160 USPQ 237, 244 (CCPA 1969) (references teach away from combination if combination produces seemingly inoperative device); see also In re Gordon, 733 F.2d 900, 902, 221 USPQ 1125, 1127 (Fed. Cir. 1984) (inoperable modification teaches away). Here, *Meyerzon* teaches rendering an electronic document for display on a web browser at the client side. Therefore, the combination of *Meyerzon* with *Lawrence* and/or in further view of *Blumenthal* to produce the presently claimed invention where an in-memory webpage representation of a web document is rendered prior to analyzing and summarizing the in-memory webpage would produce an inoperative device. Accordingly, the combination of *Meyerzon* and *Lawrence* in further view of *Blumenthal* is improper.

Accordingly, independent claims 1, 14, and 20 distinguish over *Meyerzon* and/or *Lawrence* and/or *Blumenthal for* at least the reasons stated above. Claims 2-13, and 15-19 depend from claims 1, 14, and 20 respectively. Since dependent claims contain all the limitations of the independent claims, claims 2-13, and 15-19 distinguish over *Meyerzon* and/or *Lawrence* and/or *Blumenthal*, as well. Therefore, the rejection of claims 1-20 should be reversed.

CONCLUSION

For the reasons stated above, Appellants respectfully contend that each claim is patentable. Therefore, reversal of all rejections is courteously solicited.

Respectfully submitted,

Dated: January 17, 2006 By: 

Jon Gibbons
Registration No. 37,333
Attorney for Appellants

Fleit, Kain, Gibbons, Gutman & Bongini
One Boca Commerce Center, Suite 111
551 N.W. 77th Street
Boca Raton, FL 33487
Tel. (561) 989-9811
Fax (561) 989-9812

VIII. CLAIMS APPENDIX

1. A method for browser-enhanced web crawling associated with a network of hub processing units coupled to a plurality of information processing units over a network, the method executed by a web crawler on a hub processing unit associated with the network comprising:

retrieving a web document at an address, and extracting contents of the web document for rendering an intermediate dynamically constructed in-memory webpage representation of the web document at a hub processing unit which is formatted as if displayed for viewing on an end-user's web browser;

loading secondary documents associated with the web document in order to render the secondary documents as part of the in-memory webpage representation, wherein the secondary documents include one or more images with textual content embedded therein, wherein the hub processing unit renders the in-memory webpage prior to analyzing and summarizing the in-memory webpage;

analyzing and summarizing the in-memory webpage representation to produce a text map for the webpage document of the textual contents; and

using optical character recognition on the images to extract textual content for adding to the textual map for the webpage document.

2. The method as defined in claim 1, wherein the retrieving the web document at an address further comprises retrieving a document at an address selected from the group of addresses consisting of a nodal address, a network address, a URL and equivalents.

3. The method as defined in claim 1, wherein the one or more images with textual content embedded therein include at least one of an in-line GIF image and an in-line JPEG image.

4. The method as defined in claim 1, wherein the loading secondary documents further comprises the loading of secondary documents including one or more Java applets with textual content embedded therein.

5. The method as defined in claim 1, wherein the loading secondary documents further comprises the loading of secondary documents including web documents selected from the group of documents consisting of in-line frames, frames, and equivalents.

6. The method as defined in claim 4, wherein the loading secondary documents further comprises the loading of secondary documents including one or more Java Script components with textual content embedded therein.

7. The method as defined in claim 1, wherein the retrieving the web document further comprises performing the following sub-steps of:

initializing a first list with seed values;

checking if there are any URLs to be processed and in response that any URL exists to be processed then performing the following sub-steps of:

determining if a URL is in a second list; and in response that a URL is not in the second list then performing the following sub-steps of:

inserting the URL into the first list;

scheduling the URL for crawling;

crawling the URL when scheduled to do so;

removing the URL from the first list after the scheduled crawling;

entering the URL into the second list; and

repeating the checking step until there are no more URLs to be processed;

where if the determining step determines that the URL is in the second list then repeating the checking step until there are no more URLs to be processed.

8. The method as defined in claim 7, wherein the sub-step of initializing a first list with seed values further includes the list being a URL pool.

9. The method as defined in claim 7, wherein the sub-step of determining if a URL is in a second list further includes the second list being a visited pool.

10. The method as defined in claim 7, wherein the sub-step of crawling further comprises the sub-steps of:

issuing an HTTP command to a web server named in the URL;

receiving contents of an HTML page as a result of the issued HTTP command; and

passing on the contents of the HTML page to a Page Rendering subroutine.

11. The method as defined in claim 10, further including the sub-steps performed by the Page Rendering subroutine comprising:

receiving the contents of the HTML page in the Page Rendering subroutine;

building an in-memory representation of a Layout for the HTML page and if more data is needed to properly form the representation, then performing the sub-steps of:

requesting additional web-based information;

gathering this additional web-based information;

inserting any URLs associated with this additional web-based information into the second list and a URL cache;

building a final amended representation; and

forwarding the final amended representation to an Extraction subroutine;

wherein, if no more data is needed to properly form the in-memory representation, then forwarding the in-memory representation to the Extraction subroutine.

12. The method as defined in claim 11, further including the sub-steps performed by the Page Extraction subroutine comprising:

accessing a set of memory structures of the Page Renderer;

copying a text portion of the structures into a text map;

inspecting any in-line GIF and JPEG image references in the memory structures;

extracting alternate text attributes;

adding the alternate text attributes to a text map;

invoking an optical character recognition engine;

analyzing any in-line GIF and JPEG images using the optical character recognition engine for text content;

extracting text content from the GIF and JPEG images;

adding text content from the images to the text map; and

forwarding the text map to a Page Summarizer subroutine.

13. The method as defined in claim 12, further including the sub-steps performed by the Page Summarizer subroutine comprising:

receiving a text map from the Page Extractor subroutine;

processing the text map in an application-specific manner;

applying data extraction patterns to the text map;

translating resultant data from the applying step;

forwarding any URLs present in the text map to a manager subroutine; and

forwarding any extracted data and metadata to application logic.

14. A computer readable medium including programming instructions, the programming instructions including instructions for browser-enhanced web crawling associated with a network of hub processing units coupled to a plurality of information processing units over a network, the browser enhanced web crawling instructions on the computer readable medium comprising:

retrieving instructions for retrieving a web document at an address, and extracting contents of the web document for rendering an intermediate dynamically constructed in-memory webpage representation of the web document at a hub processing unit which is formatted as if displayed for viewing on an end-user's web browser;

loading instructions for loading secondary documents associated with the web document in order to render the secondary documents as part of the in-memory webpage representation, wherein the secondary documents include one or more images with textual content embedded therein, and wherein the hub processing unit renders the in-memory webpage prior to analyzing and summarizing the in-memory webpage;

analyzing and summarizing instructions for analyzing and summarizing the in-memory webpage representation to produce a text map for the webpage document of the textual contents therein; and

using optical character recognition on the images to extract textual content for adding to the textual map for the webpage document.

15. The computer readable medium as defined in claim 14, wherein the retrieving instructions for retrieving a web document at an address further comprises retrieving instructions for retrieving a document at an address selected from the group of addresses consisting of a nodal address, a network address, a URL and equivalents.

16. The computer readable medium as defined in claim 14, wherein the one or more images with textual content embedded therein include at least one of an in-line GIF image and an in-line JPEG image.

17. The computer readable medium as defined in claim 14, wherein the loading secondary documents further comprises the loading of secondary documents including one or more Java applets with textual content embedded therein.

18. The computer readable medium as defined in claim 14, wherein the loading instructions for loading secondary documents further comprises loading instructions for loading of secondary documents including web documents selected from the group of documents consisting of in-line frames, frames, and equivalents.

19. The computer readable medium as defined in claim 17, wherein loading secondary documents further comprises the loading of secondary documents including one or more Java Script components with textual content embedded therein.

20. A browser-enhanced web crawling unit associated with a network of a plurality of hub processing units coupled to a plurality of information processing units over a network, the browser enhanced web crawling unit on a hub processing unit comprising:

a retrieval unit for retrieving a web document at an address, and extracting contents of the web document for rendering an intermediate dynamically constructed in-memory webpage representation of the web document at a hub processing unit which is formatted as if displayed for viewing on an end-user's web browser;

a loader for loading secondary documents as required associated with the web document in order to render the secondary documents as part of the in-memory webpage representation, wherein the secondary documents include one or more images with textual content embedded therein, wherein the hub processing unit renders the in-memory webpage prior to analyzing and summarizing the in-memory webpage;

a summarizer for analyzing and summarizing the in-memory webpage representation to produce a text map for the webpage document of the textual contents therein; and

an optical character recognition engine for use on the images to extract textual content for adding to the textual map for the webpage document.

IX. EVIDENCE APPENDIX

NONE

X. RELATED PROCEEDINGS APPENDIX

NONE